Depth Estimation of a Single RGB Image with Semi-supervised Two-stage Regression

Jun Chi
Ocean University of China
Songling Road. 238
Qingdao, Shandong, China
+86-15963247990
chijun4766@stu.ouc.edu.cn

Jie Gao
Ocean University of China
Songling Road. 238
Qingdao, Shandong, China
+86-17860765210
Gaojie9186@stu.ouc.edu.cn

Lin Qi
Ocean University of China
Songling Road. 238
Qingdao, Shandong, China
+86-18661886965
qilin@ouc.edu.cn

Shu Zhang
Ocean University of China
Songling Road. 238
Qingdao, Shandong, China
+86-18560612449
zhangshu@stu.ouc.edu.cn

Junyu Dong*
Ocean University of China
Songling Road. 238
Qingdao, Shandong, China
+86-13708980025
dongjunyu@ouc.edu.cn

Hui Yu
University of Portsmouth
Winston Churchill Avenue
Portsmouth, United Kingdom
+86-16678985286
hui.yu@port.ac.uk

ABSTRACT

Obtaining accurate depth estimation at low computational cost is a major problem in the field of computer vision. To tackle this problem, we propose a framework that integrates different neural networks, for predicting the corresponding depth from a single RGB image and sparse depth samples. This method combines two different types of deep learning frameworks with the best performance, including the improved Residual Neural Network and conditional generation adversarial network (cGAN). It has been proved that the improved ResNet has strong depth prediction capability, but the depth map is still incomplete in detail. We improve the existing cGAN model to enhance ResNet-based depth prediction. Experiments compared with stage-of-the-art are performed on publicly available data sets. And the results demonstrate that the proposed two-stage deep regression model is superior to other existing methods of the same type.

CCS Concepts

CCS → Computing methodologies → Computer graphics → Image manipulation → Image processing.

Keywords

Depth prediction; a single RGB image; sparse depth samples; neural networks.

1. INTRODUCTION

Depth prediction plays an important role in industrial applications such as augmented reality (AR), 3D mapping, robotics and autonomous driving. Due to the high system cost and limited operating conditions, traditional methods of depth capturing are difficult to use for most cases. For example, based on laser or depth sensor methods, 3D Lidar is over-costly and provides only sparse measurements for distant objects. Depth sensors based on structured light cannot perform well under high sunlight conditions and can only measure depth from close objects. However, using a computational model to estimate depth from a single RGB image provides a more feasible and preferred way to capture a depth map at a lower cost and wider operating conditions. Recently, although research work over the past decade has focused on RGB-based depth prediction, including deep learning-based prediction methods, the reliability and accuracy of these methods are still far from practical. For example, the most advanced RGB-based depth prediction method [1-3] produces a result with an average error of more than 50 cm (calculated by root mean square error) in an indoor scene (using the NYU-v2 data set [4]). This method performs even worse outdoors, producing an average error of at least 4 meters on the Make3D dataset [5] and an error of at least 7 meters on the Kitti dataset [6]. From the experimental results, it can be seen that using a computational model to estimate the depth from a single RGB image is a challenging task, as a result we add a constraint, a set of sparse depth samples to improve the accuracy of the calculation model. Currently, regression networks are used to predict the depth of a single RGB image, and typically only a single deep learning model is used. We obtain dense depth predictions from a single RGB dense image and a set of sparse depth samples based an existing deep regression model. Our method significantly reduce the potential limitations of RGB-based depth estimation. The effectiveness of using sparse depth measurements as part of the system input is demonstrated in the work of in Ma [7] et al. We use a two-stage convolutional neural network to learn the deep regression model of depth prediction. The experimental results show that by adding at least 100 depth samples and using the two-stage depth estimation model on the NYU-Depth-V2 data set, the root mean square error is reduced by more than 40%. Our results can contribute to develop the field of driverless and consumer sensors. However, there are a lot of geometric details that are not accurate enough and objects are occluded in indoor scenes.

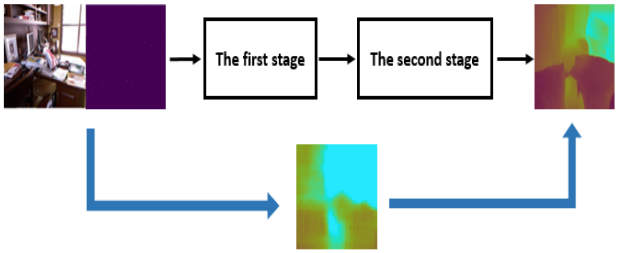

Figure 1. Proposed deep prediction framework. The RGB image and the corresponding sparse depth are used as the first-stage network input, and the first-stage network result is input as the second-stage network, and finally a more accurate depth map is obtained.

So we propose a two-stage depth estimation calculation model. The proposed model framework is shown in Figure 1. In the first stage, the improved ResNet [7] is trained to obtain a rough depth map, and the improved cGAN model [8] in the second stage is used to redraw the rough map and get the final result. Accurate depth information is important for many computer vision tasks and engineering development. The main contribution of this paper is that we propose a two-stage deep regression model, and use sparse samples and RGB images as input to the regression model, and predict the end-to-end depth image. The prediction accuracy of our method is significantly improved than the most advanced method of this type.

2. RELATED WORK

2.1 Depth Prediction

Depth prediction based on a single RGB image is getting more and more attentions from researchers, but it is still very difficult to recover detailed depth using only a single RGB image. To achieve this task, the deep learning method has been successfully applied to the depth prediction problem over the past years. For example, convolutional neural networks and their variants have brought the dawn of challenging deep prediction tasks and yield the best results. Eigen et al. [9] proposed a two-stack convolutional neural network (CNN), one predicting the global coarse scale and the other predicting the local details. Recently, the residual network (ResNet) has successfully solved the problem of gradient disappearance in deeper networks.

These methods have also been applied to the depth prediction of a single image. Laina et al. [3] developed a deep full convolution residual network based on ResNet, the upsampling block is designed not only to obtain high resolution depth but also achieves higher precision than [11]. In addition, Roy et al. [12] tackled the depth estimation by integrating random regression forest and convolutional neural networks and introduced an end-to-end architecture. Ma et al. [13] developed a deep regression model to learn direct mapping from sparse depth to dense depth. The framework was a self-supervised training framework that requires only a series of RGB images and sparse depth images to perform deep prediction tasks.

2.2 Reconstruction of Sparse Depth Data

The reconstruction of sparse depth data mainly includes two methods. One is to use sparse representation of depth signals. For example, Hawe et al. [14] assumed that the disparity map is sparse on the wavelet basis, and reconstructed the dense disparity image using the conjugate subgradient method. The second is to use sensor fusion techniques to try to improve depth prediction by fusing additional information from different sensor modalities. For example, Liao et al. [15] proposed a CNN that uses RGB images and optical flow images as inputs to predict depth. In contrast, inspired by the residual network (ResNet) successfully solving the gradient disappearance problem when training deep networks based on the work of [7] et al., we use RGB images and corresponding sparse depth samples as the input of the first stage network, and it learns better cross-modal representations for RGB and sparse depth and achieves significantly higher accuracy. Furthermore, the two stage regression model predicts a full resolution depth image.

2.3 Generative Adversarial Networks

In recent years, generative adversarial networks (GANs) and their variants have attracted many researchers' attentions. GANs have achieved great success in many fields, such as face image generation task [16], image-to-image conversion [8]. Others also use GAN for image-to-image mapping, unconditionally, which has made remarkable achievements in super-resolution [17], repair [18]. Some researchers also try to use GANs to predict depth maps from single RGB images and sparse depth samples. For example, Hyungjoo et al. [19] designed a full-convolution multi-scale network based on GANs model to optimize the estimated depth maps.

Recently, most of the depth estimation methods improve the network prediction ability by increasing the number of layers or changing the network structure. But they require a lot of training data to achieve significant improvement. For the multi-stage regression network model, there is still room for improvement. In order to reduce the complexity of the network with a small amount of training data, we propose a two-stage depth regression model, which integrates two different types of regression network models for the task of depth prediction, they are semi-supervised improved ResNet and cGAN models. The two-stage network model balances the high cost and speed requirements. It uses the excellent performance of cGAN in the regression task to further optimize the preliminary depth prediction results. As far as we know, our work is the first time to integrate semi-supervised network with conditional generation antagonism network (cGAN) for depth prediction.

3. METHOD

In this section, the proposed two-stage depth prediction network will be introduced in detail. The main purpose of the proposed method is to predict a depth map from a single RGB image and a corresponding sparse depth. A two-stage depth regression model is proposed to achieve this good. The model combines the advantages of two networks. The first-stage network adds sparse depth samples as a priori to improve accuracy; the second-stage network performs well on regression tasks to optimize the results of deep prediction. In the first stage, an improved ResNet network [7] is used to obtain a rough initial depth prediction. In the second stage, we propose a new cGAN to optimize the first stage depth map for a refined result.

3.1 Network Architecture

The network architecture of our two-stage deep regression model is shown in Figure 2. In the first stage, an improved ResNet-50 model is used to obtain the initial depth map [7]. In the second stage, an improved cGAN model based on the existing cGAN model [8] is proposed to optimize the depth map the first stage. The generator of the cGAN model uses the ‘U-Net’network architecture.

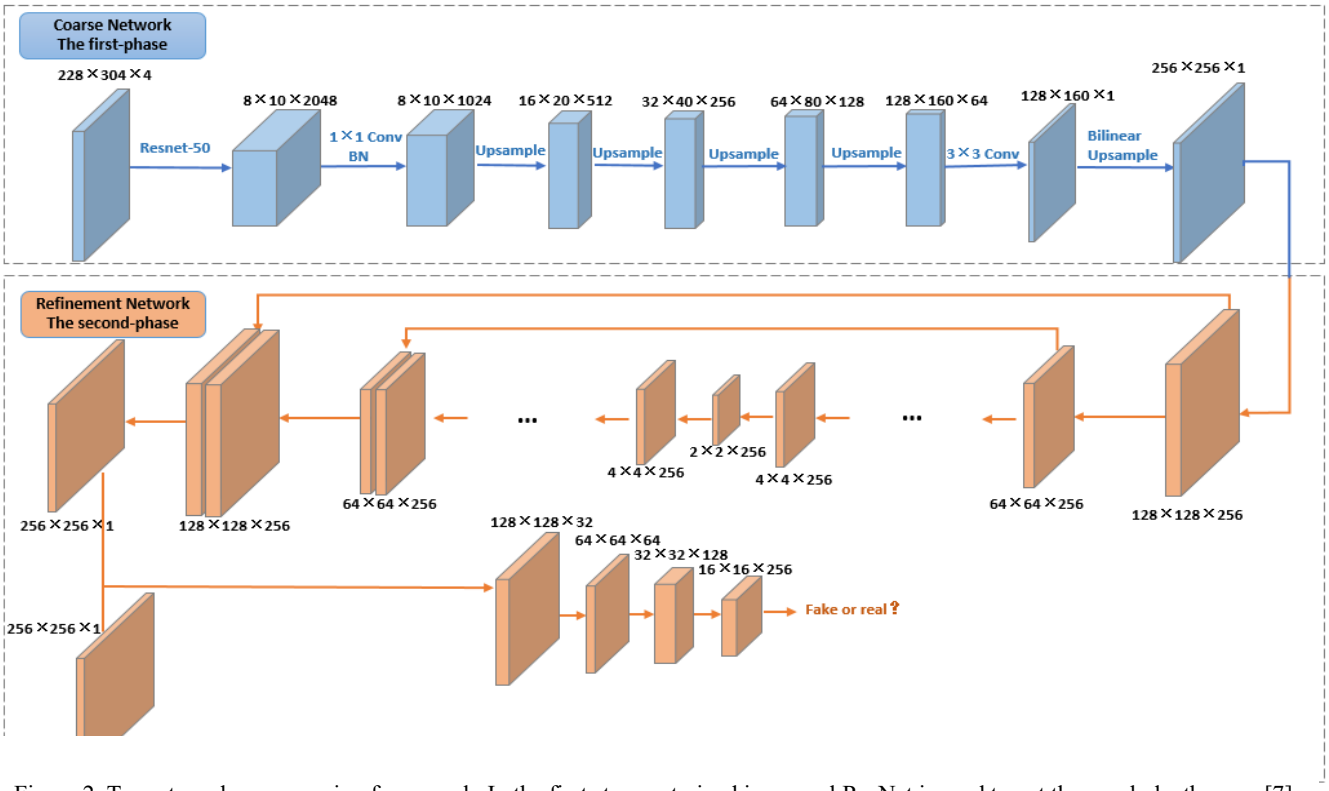

Figure 2. Two-stage deep regression framework. In the first stage, a trained improved ResNet is used to get the rough depth maps [7], we concatenate the three-channel RGB images and the one-channel sparse depths. The first-phase network builds upon ResNet-50, and adds some convolutions layers and upsample layers. In the second stage, we use a new cGAN based on the existing cGAN model [8]. We use the ResNet blocks replace the some middle layers in the 'U-Net' and reserves the skip connections.

3.2 ResNet

In our experiments, we found that the ResNet-based network architecture has good performance in depth prediction from single RGB image and sparse depth samples. Adding bottleneck structure (with encoder and decoder) on this basis can make the results more accurate. For the data set used in our experiments, ResNet-50 is utilized and improved on this basis. That is to say, we remove the last average pooling layer and linear conversion layer of ResNet-50, and add two convolution layers, four upper sampling layers and a bilinear interpolation layer. After we get the initial depth map, we add two convolution layers, four upper sampling layers and one bilinear interpolation layer to it. This is because the input size of the second stage network is 256 x 256.The resolution of depth map is changed to 256 x 256 by bilinear interpolation.

For the improved residual network based optimization method, the default loss function of the general regression network model is the mean square error (L2). In the work of Ma [7] et al., it is found that the L2 loss function yields a visually over-smooth boundary. Another choice is the reverse Huber (denoted as berHu) [20], defined as

$$B(e) = \begin{cases} \frac{e^2+c^2}{2c}, & |e| > c \\ |e|, & |e| \le c \end{cases} \quad (1)$$

As shown in the Eq. (1) above, berHu loss is equal to L1 (e) = | e | when e <[-c, c], and equal to L2 when | e | > c. Parameter c is computed as 20% of the maximum absolute error over all piexls in a batch.

3.3 cGAN

We use the improved cGAN model to fine tune and optimize the initial depth map from the first stag. Specifically, we use the improved cGAN model to tackle the transformation task from a rough depth map x to a real depth map y. It using the generator and discriminator structure in [8] and modifying them to produce better result. Both the generator and the discriminator use a module in the form of a convolution-BatchNorm-ReLu sequence. We let Ck denote a Convolution-BatchNorm-ReLU layer with k filters, and CDk denotes a Convolution-BatchNormDropout-ReLU layer with a dropout rate of 50%. The cGAN learns a mapping from the observation image x and the random noise vector z to y, and it is denoted as G: {x, z} → y. The goal of cGAN can be expressed as solving a minimum-maximum optimization problem, defined as

$$G^* = \min_G \max_D \left[E_{x,y}[\log D(x,y)] + E_{x,z}\left[\log\left(1 - D\left(x, G(x,z)\right)\right)\right] \right] + \lambda E_{x,y,z}[\log ||y - G(x,z)||_1] \quad (2)$$

where G(·) represents the generator network, D(·) represents the discriminator network, and λ represents the weighting factor. The first term of the formula represents the loss of resistance, and the second term represents the reconstruction loss at the pixel level.

In the improved cGAN model, the generator network G is trained by the Eq. (2) to spoof the discriminator network so that the depth map generated by the generator network is close enough to the true depth data. Many previous solutions for image-to-image translation tasks use an encoder-decoder network where the input goes through a series of downsampling layers until the bottleneck layer, after which the process reverses. For the mapping problem from the rough depth map to a real depth map, we improve the existing cGAN model. The generator is still similar to the original cGAN model, adding jump links according to the shape of 'U-Net' [8] low-level information can be direct transmission in the network. Additionally, we add a skip connection between each layer i and n-i, where n is the total number of layers (i =1, 2). We replace the third layer to n-3 with nine resblock Layers. The purpose of this is to increase the depth of U-Net. It solves the degradation problem while reducing the gradient disappearance. In particular, in order to make the input and output image resolution constant, we add padding to the feature map generated in the process. As for discriminator, we use an architecture named PatchGAN that only penalizes structure at the scale of patches.

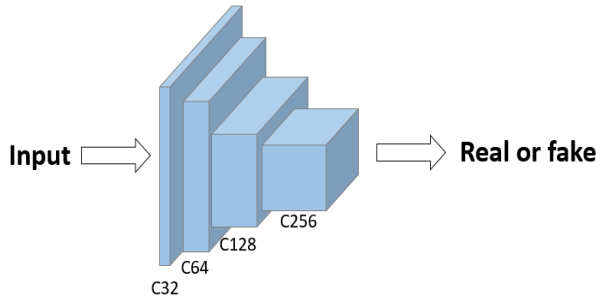

Figure 3. The encoder-decoder architecture of discriminator.

The discriminator classifies whether a N*N patch in the image is true or false. In our work we use a 70*70 patch for this process[21]. The discriminator architecture is shown in Figure 3. It should be noted that all Relus in the experiments are leaky Relus.

4. EXPERIENCES

4.1 Dataset and Error Metrics

We used Torch [22] to implement the two-stage regression network. In order to verify the validity of the proposed two-stage regression model, we performed experiments on the official indoor dataset (NYU Depth v2). 249 scenarios in the official data set were used for training and the remaining 215 were used for testing. In particular, the sparse depth samples in the first stage were derived from random sampling of ground truth depth images G*. In order to make sparse data samples valid, we use Bernoulli probability p = $\frac{n}{m}$, where m is the total number of valid depth pixels in G*. We set the pixel outside the probability p to 0, others is unchanged. We use the following evaluation indicators to quantitatively evaluate the predictions:

- RMSE: $\sqrt{\frac{1}{N}\sum_{y_i\in|N|}|y_i-y_i^*|^2}$ (3)
- REL: $\frac{1}{N}\sum_{y_i\in|N|}\frac{|y_i-y_i^*|}{y_i^*}$ (4)
- Average $\log_{10}$ error :

$$\log_{10} = \frac{1}{N}\sum_{y_i\in|N|}|\log_{10}(y_i)-\log_{10}(y_i^*)| \quad (5)$$

- Threshold δ_i (The percentage of predicted pixels whose relative error is within the threshold.):

$$\delta_i = \frac{card(\{y_i^*:max\{\frac{y_i^*}{y_i},\frac{y_i}{y_i^*}\}<1.25^i\})}{card(\{y_i\})} \quad (6)$$

where N represents the total number of pixels in a data sample, y_i represents the depth value of the model estimate, and y_i* is the depth value in the corresponding ground truth depth data. And card is the cardinality of a set. A higher δ_i indicates better prediction.

4.2 Experiences and Details

In the first stage of our proposed two-stage regression prediction model, the output of the improved ResNet resolution is 160×128. To meet the second stage input requirement, we add a bilinear upsampling layer to change it to 256×256. The weight of ResNet in the coding layer is initialized by using a pre-trained model on the ImageNet data set [23]. We use a small batch size of 10 and train it for 20 epochs. The learning rate starts at 0.01, and is reduced to 20% every 5 epochs. In the second stage, the proposed batch size of the new cGAN model is 1 with 200 epochs training. We adopt the minibatch SGD and Adam solution [24]. The learning rate is 0.0002, and the momentum parameters are β_1=0.5, β_2=0.999. The quantitative comparisons of NYU depth v2 are shown in Table 1.

Table 1. Comparison of depth prediction results on NYU depth-v2 [4]

Method	RMSE	REL	log 10	δ_1	δ_2	δ_3
Karsh et al.[26]	1.12	0.374	0.134	—	—	—
Li et al.[25]	0.821	0.232	0.094	0.621	0.886	0.968
Liu et al.[1]	0.824	0.230	0.095	0.614	0.883	0.971
Roy and Todorovic[12]	0.744	0.187	0.078	—	—	—
Eigen and Fergus[3]	0.641	0.158	—	0.769	0.950	0.988
Liao et al-225[10]	0.442	0.104	—	0.878	0.964	0.989
Ma et al-200[8]	0.230	0.044	—	0.971	0.994	0.998
ours	**0.128**	**0.041**	**0.081**	**0.935**	**0.983**	**0.998**

As observed from the comparison results, our model achieves most satisfiy results. Especially, the rms error is improved by 40% compared to the same depth prediction using RGB and sparse depth samples. In addition, the depth map visually results predicted by the two-stage regression model are shown in Figure 4.

Figure 4. Predictions on NYU-Depth-v2. From left to right: RGB images, ground truth, the results of Ma et al. [7] and our results.

5. CONCLUSIONS

We propose a semi-supervised two-stage regression model to perform the task of predicting depth from a single RGB image and sparse depth samples. It turns out that a two-stage regression model can provide better performance than a one-stage model. Future work we will focus on how to improve the accuracy of the details.

6. REFERENCES

[1] F. Liu, C. Shen, and G. Lin, "Deep convolutional neural fields for depth estimation from a single image," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 5162–5170.

[2] D. Eigen and R. Fergus, "Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture," in Proceedings of the IEEE

[3] I. Laina, C. Rupprecht et al., "Deeper depth prediction with fully convolutional residual networks," in 3D Vision (3DV), IEEE 4th International Conference, 2016, pp. 239–248

[4] N. Silberman, D. Hoiem et al., "Indoor segmentation and support inference from rgbd images," Computer Vision–ECCV 2012, pp. 746–760, 2012.

[5] Ashutosh Saxena, Min Sun, and Andrew Y. Ng. Make3d: Learning 3d scene structure from a single still image. IEEE Transactions on Pattern Analysis and Machine Intelligence 31(5):824–840, 2009.

[6] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? the kitti vision benchmark suite," in Conference on Computer Vision and Pattern Recognition (CVPR), 2012.

[7] Ma F, Karaman S. Sparse-to-Dense: Depth Prediction from Sparse Depth Samples and a Single Image [J]. 2017:1-8.

[8] P. Isola, J. Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," pp. 5967–5976, 2016.

[9] D. Eigen, C. Puhrsch, and R. Fergus, "Depth map prediction from a single image using a multi-scale deep network," in Advances in neural information processing systems, 2014, pp. 2366–2374.

[10] Y. Liao, L. Huang et al., "Parse geometry from a line: Monocular depth estimation with partial laser observation," in Robotics and Automation (ICRA), 2017 IEEE International Conference on. IEEE, 2017, pp. 5059–5066.

[11] D. Eigen and R. Fergus, "Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture," in Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 2650–2658.

[12] A. Roy and S. Todorovic, "Monocular depth estimation using neural regression forest," in Computer Vision and Pattern Recognition, pp. 5506–5514, 2016.

[13] Ma F, Cavalheiro G V, Karaman S. Self-supervised Sparse-to-Dense: Self-supervised Depth Completion from LiDAR and Monocular Camera [J]. 2018.

[14] S. Hawe, M. Kleinsteuber, and K. Diepold, "Dense disparity maps from sparse disparity measurements," in Computer Vision (ICCV), 2011 IEEE International Conference on. IEEE, 2011, pp. 2126–2133.

[15] M. Mancini, G. Costante et al., "Fast robust monocular depth estimation for obstacle detection with fully convolutional networks," IEEE/RSJ International Conference. Intelligent Robots and Systems, 2016, pp. 4296–4303.

[16] Choe J, Park S, Kim K, et al. Face Generation for Low-Shot Learning Using Generative Adversarial Networks. 2017 IEEE International Conference on Computer Vision Workshops (ICCVW), 2017:1940-1948.

[17] Ledig C, Theis L, Huszar F, et al. Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network. IEEE Conference on Computer Vision and Pattern Recognition 2016.

[18] Pathak D, Krahenbuhl P, Donahue J, et al. Context Encoders: Feature Learning by Inpainting [J]. 2016.

[19] Hyungjoo Jung, Youngjung Kim, Dongbo Min, and Kwanghoon Sohn. Depth prediction from a single image with conditional adversarial networks. In IEEE International Conference on Image Processing, pages 1717–1721, 2017.

[20] Laurent Zwald and Sophie Lambertlacroix. The berhu penalty and the grouped effect. Statistics, 2012.

[21] Phillip Isola, Jun Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. Pages 5967–5976, 2016.

[22] R. Collobert, K. Kavukcuoglu, and C. Farabet, "Torch7: A matlab-like environment for machine learning," in BigLearn, NIPS Workshop, 2011.

[23] O. Russakovsky, J. Deng et al., "Imagenet large scale visual recognition challenge," International Journal of Computer Vision, vol. 115, no. 3, pp. 211–252, 2015.

[24] M. Mathieu, C. Couprie, and Y. LeCun. Deep multi-scale video prediction beyond mean square error. ICLR, 2016. 2, 3

[25] Bo Li, Chunhua Shen, Yuchao Dai, Anton Van Den Hengel, and Mingyi He. Depth and surface normal estimation from monocular images using regression on deep features and hierarchical crfs. In IEEE Conference on Computer Vision and Pattern Recognition, pages 1119–1127, 2015.

[26] Kevin Karsch, Ce Liu, and Sing Bing Kang. Depth extraction from video using non-parametric sampling. In European Conference on Computer Vision, pages 775–788, 2012.